AI Ethics

4-20-16
Some ethical issues for AI researchers

**Short term**
- Economic upheaval
- Misuse of technology
- Assigning responsibility

**Long term**
- Autonomy
- Paperclip world
- Singularity
AI displacing jobs
2010 flash crash
Military robotics
Autonomous cars
Discussion questions

What are the redistributive consequences of AI? Are they fundamentally different from other technological innovations?

How should autonomous vehicles solve the trolley problem? Does this even matter if autonomous vehicles are safer than human drivers?

Who should be blamed if an autonomous agent causes harm? Ethically? Legally?

What safeguards can/should AI developers put in place to ensure that the technology isn’t misused?
Single-task optimizers
Singularity
Making true AI safe

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
Discussion questions

How can we implement single-task agents to ensure they are safe?

Is the singularity possible? Are we close to it? Should we fear the singularity or hope for it?

What does provably safe AI mean? How can it be achieved? Would it look more like Russell’s air traffic control analogy or Asimov’s laws of robotics?